Comparing predictive machine learning algorithms in fit for work occupational health assessments

Charapaqui-Miranda, Saul[1][0000-0002-5562-0464], Arapa-Apaza, Katherine[1][0000-0002-8640-2515], Meza-Rodriguez, Moises[1][0000-0002-5806-9014], and Chacon-Torrico, Horacio[1][0000-0003-4573-2099]

Universidad Peruana Cayetano Heredia, Lima, Peru

Abstract. Some studies have tried to develop predictors for fitness for work (FFW). This study assessed the question whether factors used in the occupational medical practice could predict an individual fit for work result. We used a Peruvian occupational medical examination dataset of 33347 participants. We obtained a reduced dataset of 2650. It was split into two subsets, a training dataset and a test dataset. Using the training dataset, logistic regression , decision tree, random forest, and support vector machine models were fitted, and important variables of each model were identified. Hyperparameter tuning was an important part in these non-parametric models. Also, the Area Under the Curve (AUC) metric was used for Model Selection with a 5-fold cross validation approach. The results shows the Logistic Regression as the most powerful predictor (AUC = 60.44%, Accuracy = 68.05%). It is important to notice the best variables analysis in fitness to work evaluation by a Random Forest approach. Thus, the best model was logistic regression. This also reveals that the criteria associated with the workplace and occupational clinical criteria have a low level of prediction. Further studies should be done with imbalanced data to process bigger datasets, in consequence to obtain more robust models.

Keywords: Machine Learning · Occupational Health · Data Science

1 Introduction

The role of Big data and machine learning in health care is an emerging area that enables the prediction and understanding of health determinants all along the continuum of care [1]. The main driver that has consistently facilitated its ever-increasing implementation is the ability to analyze huge volumes and varieties of structured and unstructured data not previously feasible. Epidemiological surveillance, signal detection and disease modelling are among the many examples researchers have developed [2]. Predictive algorithms are becoming transcendent in public health and epidemiology. Moreover, the transition from paper medical records to Electronic Health Records (EHR) has led to an exponential growth of data collection. As a result, big data provides wonderful opportunities for physicians, epidemiologists, and health policy experts to enhance data driven decisions that will ultimately affect health [3].

Occupational health aims to promote and maintain workers with the highest degree of physical, mental and social well-being. The prevention and control of occupational diseases has recently become a relevant problem. Globally, two million people die each year because of occupational accidents and work related injuries and illnesses [4, 5]. Annually, work related diseases are estimated to occur in 160 million people, and approximately 58 million of them are to be absent during four workdays a year. Nonetheless, workplace accidents and injuries can be supervised and prevented with effective occupational health and safety policies [6]. The assessment of fitness for work (FFW) evaluates whether an individual is fit to execute his or her work tasks without risk to self or others [7]. These evaluation notes are documents issued by doctors that state the physical and mental fitness of an individual for work [8]. Thus, knowledge of both labour and health determinants is required to assess FFW evaluations. These documents are the cornerstone of occupational health assessment, for which patients have to surpass several general practitioners (GP), audiometric, ophthalmic, psychological and cardiological evaluations. Even when there are several guidelines and protocols for the FFW evaluation [9], reports indicate that physicians tend to rate heterogeneously these assessments [10].

The present study aims to develop and test machine learning models for FFW in an Peruvian urban occupational health medical evaluation dataset and compare their performance using several metrics. The best models obtained could be implemented in the occupational assessment as an aid for the physicians.

2 Methods

2.1 The dataset

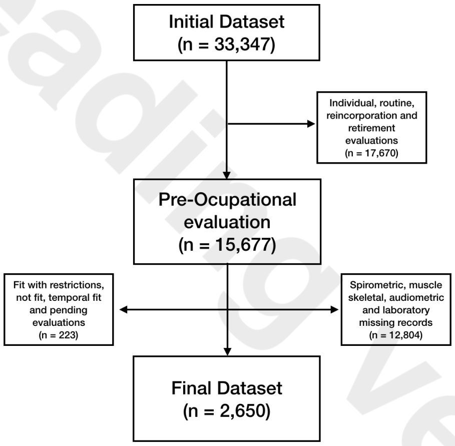
A comprehensive occupational health assessment raw dataset was prepared, analyzed and modelled for the algorithm construction. The data records come from an urban occupational medicine private clinic in Lima, Peru. Two excel files corresponding to the 2017 and 2018 data, with anonymized records and respective permissions, were physically delivered to the authors. This clinic provides occupational health consultations to third party businesses and firms that need their employees medically evaluated. Delivered data sets included all kind consultations and diagnosis registered during the aforementioned period. All recorded participants were adults between 18 and 88 years. The analyzed dataset is not publicly available.

Merged databases had 33,347 observations with 221 variables. Since the dataset included all types of consultations and results, most of the included fields contained missing values. Among the recorded categories in the dataset sociodemographics, consultation characteristics, laboratory and clinical evaluation fields were present. Most of the variables contained unstructured text including diagnosis, medical notes and recommendations. Only 14 variables contained no missing values and only 50 (22.6%) out of the 221 variables had 15% or less missing values.

2.2 Data preparation

We choose work-entry occupational health assessment for the general model (n=15,677) because the aim of this examination is to screen job applicants who may have an increase risk for occupational disease or injury in their work. Likewise many employers and other stakeholders believe that examinations prevent occupational diseases an sickness absence [11]. Considering an occupational health core evaluation that includes spirometric, muscle skeletal, audiometric and laboratory assessments, we dropped records that had missing values in these key components (see figure 1).

Fig. 1. Dataset preparation and record selection


Initial Dataset (n = 33,347)
Individual, routine, reincorporation and retirement evaluations (n = 17,670)
Pre-Ocupational evaluation (n = 15,677)
Fit with restrictions, not fit, temporal fit and pending evaluations (n = 223)
Spirometric, muscle skeletal, audiometric and laboratory missing records (n = 12,804)
Final Dataset (n = 2,650)

2.3 Feature selection

Even when the amount of variables provided in the dataset was large, the final features used for the predictive model was dramatically reduced. After we reduced the dataset we selected features that had no missing values. The final step used a heuristic feature selection where we kept the following variables:

"year", "id", "sex", "age", "job_type", "right_audiometry", "left_audiometry", "hearingloss_bilateral", "spirometry", "bmi", "hemoglobyn", "glucose","bioq_cholesterol_total", "musculoskeletal"

Then for the purpose of the model construction we converted the columns that contained categorical values to numerical values by means of the One-hot-encoding algorithm.

2.4 The outcome

A comprehensive systematic review of literature regarding FFW assessments described that less than 0.6% of the patients had their assessment as medical unfit for their duties [9]. Hence, we opted to select and filter (see figure 1) only two types of fitness for work status: Fit for work and Fit with recommendations. These dichotomization had a balanced distribution with 33.2% (n = 882) and 66.8% (n = 1768) fit and fit with recommendation results respectively.

2.5 Data analysis

The correlation analysis between the data of the quantitative inputs (age, BMI, hemoglobin, glucose, cholesterol) showed no strong relationship between them. The final dataset (n=2650) consisted on 14 features and the FFW outcome that is used to feed machine learning models such as Logistic Regression, Support Vector Machine (SVM), Random Forest and Decision Tree. We chose non-parametric methods because they presuppose less assumptions of the data distribution than parametric methods. We applied a 5-fold cross validation technique.

2.6 5-fold cross validation

There are several methods to diagnose model performance but for our specific type of data, 5-fold cross validation was the best choice to avoid overlapping data processing and to diminish the pessimistic bias [12].

We had to split the original dataset in training and test data to fit the model and evaluate the performance with unseen data, respectively. A good way to obtain a generalized model is to split it in training (70%) and testing (30%) portions. Subsequently, the training data is split in training and validation data according to the number of folds (k = 5) to evaluate the performance of each model. At last, the performance is processed by an arithmetic mean of each fold to obtain a robust metric such as Accuracy or Area Under the Curve (AUC) from the Receiver Operating Characteristics (ROC) curve to compare within models. The performance of non-parametric models depends on different parameter values called hyperparameters [13]. We can take advantage of cross validation to compare those with different methods. The hyperparameter tuning include C, kernel and gamma for Support Vector Classifier, max_depth, min_samples_split, min_samples_leaf d max_features for Decision Tree, Grid layout and Random layout in logistic regression.

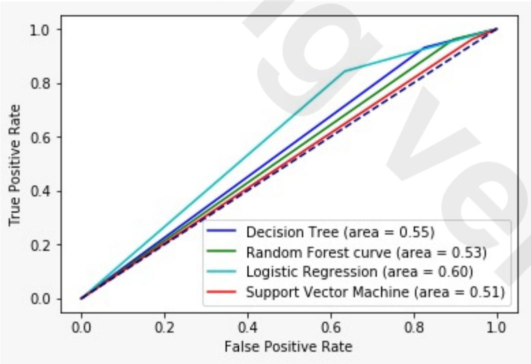
3 Results

The final dataset was analyzed and processed using a Jupyter Notebook with Python version 3. The Scikit-Learn library provide machine learning models and the function to split the data in training and test data. The function GridSearchCV of the library Scikit-Learn saved us a lot of time on the model fitting and hyperparameter tuning. GridSearchCV contains these powerful features like k-fold cross-validation based on metrics evaluations and hyperparameter tuning. The best hyperparameters chosen are shown in the Table 1 [14]. Table 2 shows performance for each machine learning models based on accuracy and AUC-ROC, but the final evaluation for model selection to chose the best generalize model is the Logistic Regression with 60.44%. As it is shown graphically the AUC of each ROC curve in the Figure 2.

Table 1. Machine learning models and its parameters

Machine Learning Model	Hyper Parameters
Decision Tree	’max depth’:7, ’min samples split’=350
Random Forest	’max depth’:7, estimators = 500, ’criterion’:’entropy’
Logistic Regression	’C’: 1, ’penalty’: l1
Support Vector Machine	’C’: 10, ’gamma’: 0.001

Fig. 2. ML models metrics evaluation



The hyper-parameter optimization of the models were selected by the Grid Search technique and the learning process was done by a k-fold cross validation approach [13].Table 2 shows the Accuracy and the Area Under the Curve (AUC) evaluation metrics from the test dataset with its respective machine learning models. Those results are compared according to its metrics and the best model for FFW prediction is the Logistic Regression model.

Table 2. ML models metrics evaluation

Machine Learning Model	Accuracy	AUC-ROC
Decision Tree	67.30%	55.24%
Random Forest	66.92%	53.26%
Logistic Regression	68.05%	60.44%
Support Vector Machine	65.41%	51.04%

4 Discussion

In this case study, we used a non-parametric ensemble machine learning models for supervised learning [15]. As the problem in study was a binary classification, the outcome of this kind of models is a probability value between [0,1], then a threshold by the researcher is applied to determine the classification. The threshold value will affect greatly the final model results, so one option to surpass this kind of matter is to evaluate every threshold value in a curve called the ROC curve for every model. Evaluating ROC curve metrics has other benefits like knowing the rate of false positives and true positives. Finally, we measured the Area Under of the Curve (AUC) to quantitatively choose the best model.

We chose an evaluation metric such as the AUC because we add an evaluation metric like the area under the curve (classification problem where we depend on the probabilities of the models and a static threshold).

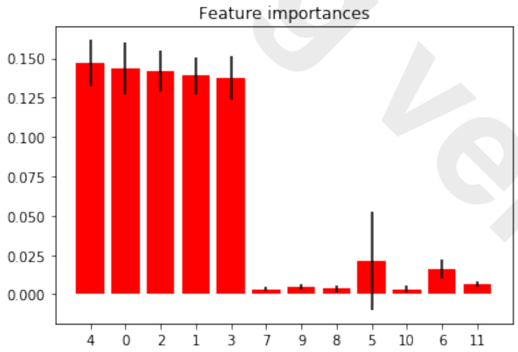


Fig. 3. ML models metrics evaluation

Fitness-to-work examination requires an objective assessment of the physical and mental health of employees. Occupational physicians have to be aware of the working conditions of characteristics of jobs, to ensure that the workers will not be a hazard to themselves or others. So there are factors like the working conditions and the health standards [16].

The variables we used to predict fitness for work were fit which meant the employee is able to perform the job without danger to self or others, without any restrictions. Likewise we considered fit subject to modifications, The reason why we used it is because we realized we had imbalanced data, (i.e. we had few unfit observations).

We analyzed occupational health work-entry assessments and their correlation with the variable fitness for work and found that it does not have a great predictive power. In future models this technique should be reassessed since this technique could encompass bias when large number of categorical and quantitative variables are found [17]. The Feature selection technique based on a Random Forest is very popular among data scientist practitioners, but the importance values of each variable are the bias when we use both numeric and categorical multi-class variables.

On the other hand, we analyzed determining health standards. We used the feature selection technique based on a random forest machine learning model technique, to predict the variables that have greater predictive power, finding that those with the highest predictive value were glucose, age, hemoglobin, BMI and cholesterol. As it is shown in Figure 3. The binary classification problem has been done correctly according to the best practices in small datasets and avoiding overlapping data issues. Although, it would be interesting to use a larger datasets with imbalanced data since some useful techniques to approach this issue like re-sampling techniques currently exist [18].

5 Conclusion

We can conclude that the best model for binary classification problem for fitness for work was the logistic regression. Likewise the variables like occupation did not had a great predictive power but further studies can improve the feature selection techniques. We recommend to do further studies in larger data using re-sampling techniques in order to enhance the performance metrics as bioinformatics data is commonly imbalanced in classification problems and, also, implement other supervised learning models like Artificial Neural Network, Bayesian algorithms, Boosting approaches, etc.

No previous study to our knowledge has tested the prediction of FFW assessment in occupational health datasets. Since routinely occupational assessments are constantly required all around the globe, we think our models and results are relevant as a novel prediction approach to this matter and that could be further developed.

References

1. Murdoch, T.B., Detsky, A.S.: The inevitable application of big data to health care. Jama **309**(13), 1351–1352 (2013)
2. Kruse, C.S., Goswamy, R., Raval, Y.J., Marawi, S.: Challenges and opportunities of big data in health care: a systematic review. JMIR medical informatics **4**(4), e38 (2016)

3. Char, D.S., Shah, N.H., Magnus, D.: Implementing machine learning in health careaddressing ethical challenges. The New England journal of medicine **378**(11), 981 (2018)
4. Mona, G.G., Chimbari, M.J., Hongoro, C.: A systematic review on occupational hazards, injuries and diseases among police officers worldwide: Policy implications for the south african police service. Journal of occupational medicine and toxicology **14**(1), 2 (2019)
5. Rommel, A., Varnaccia, G., Lahmann, N., Kottner, J., Kroll, L.E.: Occupational injuries in germany: Population-wide national survey data emphasize the importance of work-related factors. PloS one **11**(2), e0148798 (2016)
6. Saifullah, H., Li, J.: Workplace employees annual physical check-up and during hire on the job to increase health care-awareness perception to prevent diseases risk: A work for policy implementable option to global. Safety and Health at Work (2018)
7. Cox, R.A.F., Edwards, F., Palmer, K.: Fitness for Work: The Medical Aspects. Oxford University Press (2000)
8. Coggon, D., Palmer, K.T.: Assessing fitness for work and writing a fit note. Bmj **341**, c6305 (2010)
9. Serra, C., Rodriguez, M.C., Delclos, G.L., Plana, M., López, L.I.G., Benavides, F.G.: Criteria and methods used for the assessment of fitness for work: a systematic review. Occupational and environmental medicine **64**(5), 304–312 (2007)
10. Foley, M., Thorley, K., Van Hout, M.C.: Assessing fitness for work: Gps judgment making. The European journal of general practice **19**(4), 230–236 (2013)
11. Mahmud, N., Schonstein, E., Schaafsma, F., Lehtola, M.M., Fassier, J.B., Reneman, M.F., Verbeek, J.H.: Pre-employment examinations for preventing occupational injury and disease in workers. Cochrane database of systematic reviews (12) (2010)
12. Raschka, S.: Model evaluation, model selection, and algorithm selection in machine learning (2018)
13. Wong, J., Manderson, T., Abrahamowicz, M., Buckeridge, D.L., Tamblyn, R.: Can hyperparameter tuning improve the performance of a super learner? a case study. Epidemiology (Cambridge, Mass.) (2019)
14. Lee, J., Kim, H.R.: Prediction of return-to-original-work after an industrial accident using machine learning and comparison of techniques. Journal of Korean medical science **33**(19) (2018)
15. Andreas, Lindholm and Niklas, Wahlström and Fredrik, Lindsten and Thomas B., Schön: Supervised machine learning. Available at: http://www.it.uu.se/edu/course/homepage/sml/literature/lecture_notes.pdf, accessed: 2019-5-31
16. Cowell, J.: Guidelines for fitness-to-work examinations. CMAJ: Canadian Medical Association Journal **135**(9), 985 (1986)
17. Zhou, Z., Hooker, G.: Unbiased measurement of feature importance in tree-based methods. arXiv preprint arXiv:1903.05179 (2019)
18. Konno, T., Iwazume, M.: Pseudo-feature generation for imbalanced data analysis in deep learning (2018)